Cross-linguistic Readability and Controllable Difficulty: A Corpus-Based Comparison of Human and LLM Translations of Children's Literature in Romanian

Karla Csuros, Madalina Chitez, Roxana Rogobete
West University of Timisoara
{karla.csuros, madalina.chitez, roxana.rogobete}@e-uvt.ro

Abstract

Translation can systematically alter text difficulty, particularly when moving into morphologically rich languages. This study examines whether readability-constrained Large Language Models (LLMs) can mitigate difficulty shifts observed in English–Romanian translation of children's literature. We construct a paired four-condition corpus comprising English originals, published Romanian translations, readability-constrained LLM translations, and human readability adaptations (12 aligned passages; $\approx 23,000$ words). Readability is assessed using a Romanian grade-level index (LEMI) designed to be educationally comparable to Flesch-Kincaid Grade Level (FKGL), the cross-linguistic LIX metric, and morphologically informed measures derived from spaCy. Published Romanian translations are significantly more difficult than their English originals, showing higher LIX scores, grade-level estimates, and increased morphological variation. Readability-constrained LLM translation substantially reduces difficulty relative to the published versions (median $\Delta \approx 1.46$ grade levels), with significant decreases in LIX, morphological feature density, and lexical diversity (MTLD). Human adaptation yields a smaller reduction (median $\Delta \approx 0.26$). Although the direct comparison between LLM and human adaptation is marginal ($p = .055$, $r = 0.64$), LLM outputs generally produce larger reductions. These findings demonstrate that translation-induced difficulty shifts are measurable and that controllable LLM translation can modulate readability across structural, lexical, and morphological dimensions in multilingual educational contexts.

Keywords: readability assessment, translation-induced difficulty, controllable text generation

1. Introduction

Readability control is increasingly central to educational NLP, particularly in multilingual contexts where translation mediates access to age-appropriate content. While translation aims to preserve semantic content, it may inadvertently alter linguistic difficulty. In morphologically rich languages such as Romanian, translated texts can exhibit increased syntactic density, inflectional marking, and lexical formalization relative to their source texts. These shifts can raise processing demands for young readers, even when narrative content remains unchanged. Additionally, translated literature occupies a central place in the Romanian children's book market, where a substantial proportion of contemporary titles originate from English-language publishers (Sarbu, 2025; Iovanel, 2022; Cocargeanu, 2015). Translation therefore plays a key role in determining the linguistic accessibility of reading materials available to young Romanian readers. Understanding how translation influences readability is therefore both a linguistic and an educational question.

Recent advances in large language models (LLMs) enable explicit control over stylistic and structural parameters, including readability level. However, empirical evaluation of readability-controlled translation remains limited, especially for languages with high morphological complexity. In particular, it is unclear whether LLM-based translation can mitigate translation-induced difficulty shifts, and whether such mitigation operates through identifiable linguistic mechanisms.

This paper presents a controlled, within-text corpus study examining translation-induced difficulty and readability-constrained generation in Romanian children's literature. We construct a parallel four-condition corpus consisting of: (i) English originals, (ii) published Romanian translations, (iii) readability-controlled LLM translations from English into Romanian, and (iv) human readability adaptations of the published Romanian translations. Each passage ($N = 12$) is represented across conditions, enabling paired statistical comparisons.

We evaluate difficulty using both language-specific and cross-linguistic metrics, including LIX, a custom grade-level index (LEMI) for Romanian, and Flesch–Kincaid Grade Level (FKGL) for English. To move beyond surface readability measures, we additionally compute lexical diversity (MATTR, MTLD) and morphological complexity indicators derived from spaCy's Romanian large model, including average morphological feature density and inflectional diversity (wordform–lemma ratio). Statistical analyses employ paired Wilcoxon signed-rank tests with rank-biserial effect sizes.

The study addresses the following research questions:

RQ1: Do published Romanian translations ex-

Proceedings of the 2nd Workshop on Evaluating Text Difficulty in a Multilingual Context (DeTermIt! 2026), pages 1–11
11 May 2026.

hibit higher readability levels than their English source texts?

RQ2: Can readability-constrained LLM translation significantly reduce translation-induced difficulty?

RQ3: How does LLM-based readability control compare to human readability adaptation?

RQ4: Which linguistic dimensions are associated with observed difficulty shifts?

2. Theoretical Background

The intersection of translation studies, readability assessment, and generative AI provides the theoretical scaffolding for this study. We examine how translation naturally shifts text complexity, how these shifts are measured in morphologically rich languages, and how Large Language Models (LLMs) function as agents of controllable text adaptation.

2.1. Translation Universals and Complexity Shifts

Translation is rarely a neutral act of semantic transfer; it inevitably alters the stylistic and structural properties of the text. Corpus-based translation studies have long posited the existence of *translation universals*, i.e, linguistic features typical of translated texts distinct from original production (Baker, 1993). Two universals are particularly relevant to children's literature: *explicitation* and *normalization*.

Explicitation refers to the tendency of translators to spell out implicit information, often adding conjunctions and explanatory phrases to ensure clarity (Blum-Kulka et al., 1996). While intended to aid comprehension, this process frequently inflates sentence length and syntactic density. In the context of English-to-Romanian translation, this effect is compounded by systemic linguistic differences. As a Romance language with a high degree of inflection, Romanian often requires more morphological markers (e.g., definite articles attached to nouns, elaborate verbal agreement) than the analytic English structure (Pirvulescu, 2002; Ivancu, 2019; Giurgea, 2024). Consequently, a standard translation of a Grade 3 English text may inadvertently shift to a Grade 6 difficulty level in Romanian purely through obligatory grammatical expansion and the formal register often adopted by professional translators.

Cross-linguistic studies (Ciobanu et al., 2015) demonstrate that readability metrics are sensitive to the structural properties of different language families. For example, when examining texts translated into English, source language characteristics systematically influence the target text's readability profile, a phenomenon known as source language interference. Specifically, texts translated from Romance languages into English consistently yield higher Flesch-Kincaid complexity scores than original English texts or texts translated from Germanic source languages. While readability features alone may not always perfectly discriminate between original and translated texts, these metric variations underscore how systemic linguistic differences across language families can systematically impact measured text complexity during the translation process(Ciobanu et al., 2015).

2.2. Readability Assessment in Multilingual Contexts

Quantifying these shifts requires robust metrics. Traditional formulas like the Flesch-Kincaid Grade Level (FKGL) rely heavily on surface features such as sentence length and syllable count (Kincaid et al., 1975). While effective for English, these metrics often fail to capture the complexity of morphologically rich languages. The LIX (Läsbarhetsindex) metric (Björnsson, 1968) offers a more cross-linguistically valid alternative by measuring the proportion of long words (> 6 characters) rather than syllable counts, which vary wildly between Germanic and Romance languages.

However, surface metrics miss the "deep" complexity of text, such as cohesion and morphological load. Recent frameworks like ReaderBench (Dascalu et al., 2017) have moved toward multidimensional analysis, integrating lexical diversity, syntactic depth, and discourse structure. For Romanian, assessing readability requires specific attention to inflectional density, i.e., the ratio of functional morphemes to lexical roots, which significantly impacts cognitive processing load for developing readers.

2.3. Controllable Generation and Text Simplification

The advent of Large Language Models (LLMs) has introduced new paradigms for *controllable text generation*. Unlike Statistical Machine Translation (SMT), which prioritizes fidelity to the source, LLMs can be prompted to satisfy specific stylistic constraints, such as "translate for a 9-year-old" or "limit sentence complexity" (Brown et al., 2020).

Recent work in Automatic Text Simplification (ATS) suggests that LLMs can perform translation and simplification simultaneously (Martin et al., 2020). However, this *trans-adaptation* capability remains under-explored for lower-resource or morphologically complex languages. While LLMs excel at fluency, there is a risk of "stylistic flattening," where the model reduces difficulty by stripping away

the narrative voice or unique cultural markers essential to literary fiction (Maddela et al., 2021). This study addresses this gap by evaluating whether LLMs can balance the preservation of narrative intent with the structural constraints required for accessibility in Romanian.

3. Corpus and Experimental Design

3.1. Source Data

We curated a parallel corpus of 12 passages from contemporary English children's literature and their published Romanian translations. Passages were selected to represent narrative prose suitable for primary and lower-secondary readers. The mean passage length is approximately 506 words (English originals), with minor variation across conditions due to translation and editing. Each passage constitutes an independent experimental unit and is represented across multiple conditions, enabling within-text comparisons. The resulting dataset forms a small but tightly controlled parallel corpus ($\approx 23,000$ words) designed to isolate readability shifts introduced by translation and subsequent adaptation. Because the dataset contains 12 passages, the study should be interpreted as a tightly controlled pilot experiment designed to isolate translation-induced readability shifts rather than to estimate population-level parameters.

3.2. Corpus Conditions

For each passage (identified by a shared code), we constructed four aligned conditions: EN_OG (original English text), RO_OG (published Romanian translation), RO_ED (human readability adaptation of the published Romanian translation), RO_LLM (readability-constrained LLM translation from English into Romanian). This design yields a matched quartet per passage, allowing paired statistical analysis across conditions.

The RO_ED condition was created by a small readability-specialized team through manual adaptation of the published Romanian translations, without access to the original English publication. Editors were instructed to preserve narrative content and plot structure, reduce sentence length where possible, simplify lexical choices, minimize excessive formal register and complex morphological constructions,and maintain natural Romanian style. Edits were conservative and focused strictly on readability, not stylistic rewriting. The human adaptation condition was produced by a team of six editors with experience in Romanian children's literature readability. Each passage was edited independently by one editor and subsequently reviewed by another team member to ensure consistency with the readability guidelines. Disagreements regarding lexical or structural simplifications were resolved through discussion. Because the task involved controlled editing rather than categorical annotation, formal inter-annotator agreement metrics were not computed. This condition serves as a human baseline for readability-oriented adaptation.

The RO_LLM condition was generated by prompting the *ChatGPT 5.2* large language model to translate each English passage into Romanian under explicit readability constraints. The prompt specified the target grade level (identical to the original English version), preservation of narrative content and proper nouns, avoidance of unnecessary formal pronouns and politeness inflation, preference for shorter sentences and high-frequency vocabulary, and prohibition of additions, omissions, or summarization. Translations were produced in a single-pass controlled generation setup. Outputs were manually checked for content fidelity (no omissions or hallucinated content) prior to inclusion in the corpus.

3.3. Experimental Design

The study employs a within-text paired design. Each passage is compared across conditions using matched observations. This design minimizes variance introduced by genre, topic, and narrative content, and enables non-parametric paired statistical testing. Primary comparisons include: translation-induced shifts (RO_OG vs EN_OG), readability-controlled translations (RO_LLM vs RO_OG), human adaptation (RO_ED vs RO_OG), and LLM vs human adaptation (RO_ED vs RO_LLM).

Statistical analysis employs paired Wilcoxon signed-rank tests with rank-biserial effect sizes. For each comparison family (e.g., RO_ED vs RO_LLM), p-values across metrics were adjusted using the Benjamini–Hochberg procedure. The Wilcoxon signed-rank test (Taheri and Hesamian, 2013) is a non-parametric paired comparison test appropriate for small samples. Rank-biserial correlation (r) (Cureton, 1956) provides an effect-size estimate for paired ordinal comparisons. Benjamini–Hochberg correction (van Loon et al., 2017) controls the false discovery rate across multiple tests.

All preprocessing and metric computation were fully automated. The pipeline includes tokenization and morphological annotation using spaCy's RO_CORE_NEWS_LG and EN_CORE_WEB_TRF models, syllable counting via *Pyphen* (with language-specific hyphenation dictionaries), readability, lexical diversity, and morphological complexity metrics computed in Python, and statistical analysis conducted in R using the *tidyverse* and *rstatix* packages. Scripts for corpus processing, metric extraction, and statistical analysis are designed for

reproducibility and can be extended to additional passages or language pairs.

3.4. LLM Text Generation Setup

The readability-constrained translations (RO_LLM) were generated using the *GPT-5.2* large language model in a controlled, single-pass prompting setup. The model was instructed to produce Romanian translations of the English source passages while adhering to explicit readability constraints aligned with the target grade level of each text.

A standardized two-part prompting procedure was used to ensure consistency across passages. First, a fixed system-style instruction defined the model's role as a professional translator specialized in children's literature and readability control, and specified both hard constraints and soft readability guidelines. Second, for each passage, a variable prompt segment provided the target readability level and the source text. Hard constraints included preservation of semantic content (no additions, omissions, or summaries), retention of proper nouns and character names, maintenance of dialogue structure and paragraph segmentation, and avoidance of "politeness inflation" (e.g., introduction of formal pronouns or honorifics not present in the source text). The model was also instructed to preserve narrative tone and stylistic coherence.

Readability control was operationalized through soft constraints encouraging shorter sentence structures, reduced syntactic subordination, preference for high-frequency vocabulary, avoidance of nominalizations and abstract phrasing, and maintenance of clear cohesion. These constraints were framed as guidelines rather than strict rules in order to avoid distortion of meaning or narrative flow.

The prompt additionally required the model to output (i) the Romanian translation and (ii) a short list of 3–6 explicit changes applied to improve readability (e.g., sentence splitting, lexical simplification). While these notes were not included in the quantitative analysis, they were used for qualitative inspection of simplification strategies.

All translations were generated in a single pass with deterministic decoding settings (low temperature) and without post-editing. Outputs were manually checked for content fidelity (i.e., absence of omissions or hallucinated content) prior to inclusion in the corpus.

4. Metrics

We evaluate readability and linguistic complexity using a combination of established readability indices, lexical diversity measures, and morphologically informed metrics. All measures were computed automatically using a fully reproducible pipeline.

4.1. Readability metrics

For Romanian texts, we employ the **LEMI grade-level index** developed within the LEMI readability framework for Romanian children's literature (Chitez et al., 2024). The index integrates structural and lexical components, including average sentence length, proportion of complex (polysyllabic) words, proportion of unique complex word types, and overall lexical diversity. The formulation combines these components into a single grade-level estimate calibrated for Romanian educational contexts. While the specific coefficients of the LEMI index will be detailed in a forthcoming work, we utilize the metric here to provide educationally calibrated grade-level estimates distinct from raw complexity scores.

For English source texts, we compute the **Flesch–Kincaid Grade Level (FKGL)** (Kincaid et al., 1975). FKGL estimates U.S. grade-level difficulty based on sentence length and syllable counts. This metric is used primarily to contextualize the difficulty of the original English passages. The LEMI Romanian grade-level index was designed to approximate grade-level interpretation analogous to the FKGL for English. Both indices estimate school-grade readability difficulty within their respective educational systems and are calibrated to map linguistic features onto grade-level expectations. Although developed independently and trained on language-specific data, their outputs are intended to represent comparable educational grade levels rather than arbitrary complexity scores. We therefore visualize raw grade estimates on a shared axis, interpreting them as educationally aligned measures while acknowledging that linguistic scaling may differ across languages.

A potential concern is the temporal distance between the FKGL formula (1975) and the recently developed LEMI index (2024). Educational expectations, reading habits, and exposure to written language have evolved substantially over this period. In the present study, however, FKGL is used primarily as a contextual reference for the original English passages, and the LEMI index estimates Romanian grade-level readability within its contemporary educational context. To also provide a cross-linguistic readability proxy applicable to both English and Romanian, we compute the **LIX (Läsbarhetsindex)** score (Björnsson, 1968). LIX combines average sentence length and the proportion of long words (more than six characters). Its language-independent design makes it particularly suitable for comparing difficulty shifts across translation conditions.

4.2. Lexical diversity

To capture lexical variation independently of text length, we compute two widely used measures: **MATTR (Moving-Average Type–Token Ratio)** with a window size of 50 tokens (Covington and McFall, 2010), which calculates average type–token ratio across sliding windows, reducing sensitivity to overall text length, and **MTLD (Measure of Textual Lexical Diversity)** (McCarthy and Jarvis, 2010), computed using a standard threshold of 0.72. MTLD segments the text into sequential factors based on declining type–token ratio, producing a length-robust diversity estimate. Both measures provide complementary views of lexical richness and are particularly informative when evaluating simplification effects.

4.3. Morphological Complexity

To capture language-specific structural effects in Romanian, we compute morphologically informed metrics using spaCy's RO_CORE_NEWS_LG model. First, **Morphological Feature Density** is defined as the average number of morphological features assigned per token, excluding punctuation (Bentz et al., 2016). Features include grammatical categories such as case, number, gender, tense, mood, and person. This metric approximates inflectional and agreement load within the text. Morphological feature density is used here as an operational proxy for inflectional and agreement load. Additionally, we compute the **Wordform–Lemma Ratio**, as inflectional diversity is approximated by the ratio of unique surface wordforms to unique lemmas (Kettunen, 2014; Lu, 2012). Higher values indicate greater morphological variation relative to lexical base forms.

Together, these metrics allow us to assess readability shifts at multiple linguistic levels: structural (sentence length), lexical (word complexity and diversity), and morphological (inflectional density). This multidimensional approach supports a more fine-grained analysis of translation-induced difficulty and readability-controlled generation.

5. Results

All statistical comparisons use paired Wilcoxon signed-rank tests with rank-biserial correlation (r) as effect size. Reported p-values are Benjamini–Hochberg adjusted within each comparison family. Medians refer to within-passage deltas.

5.1. Translation-Induced Difficulty

We first examine whether published Romanian translations exhibit higher difficulty than their English originals.

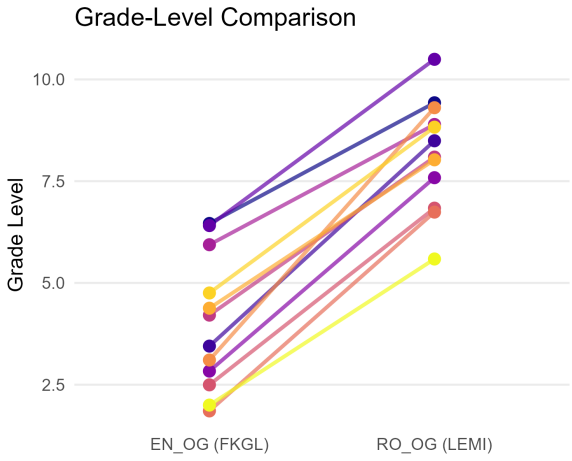

Figure 1: Comparison of English originals (FKGL) and Romanian translations (LEMI)

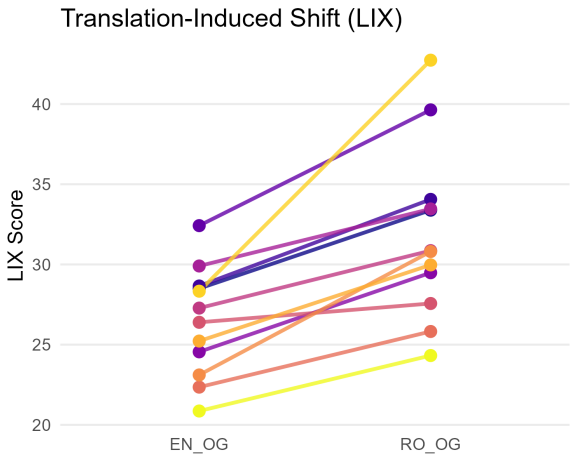

Figure 2: Cross-linguistic comparison of English originals and Romanian translations using LIX

Because FKGL and the LEMI-based index were calibrated independently, comparisons are interpreted in terms of directional and magnitude shifts in educational grade level rather than strict metric identity. As shown in Figure 1, Romanian translations (RO_OG) consistently exhibit higher grade-level estimates than the English originals (EN_OG) across passages, indicating an upward shift in educational readability level.

To provide a strictly cross-linguistic comparison, we examine *LIX* scores, which are computed using the same formula in both languages. Romanian translations show significantly higher LIX values than the English originals (median $\Delta > 0$; $p = .0025$, $r = 1.00$), reflecting increased sentence length and/or a higher proportion of long words. The consistency of this shift is illustrated in Figure 2, where nearly all passages display an upward trajectory from EN_OG to RO_OG.

Morphological complexity patterns further sup-

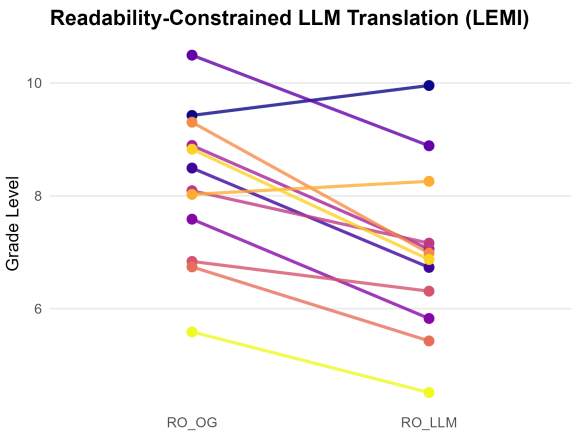

Figure 3: Paired comparison of published Romanian translations and readability-constrained LLM translations using the LEMI grade-level index

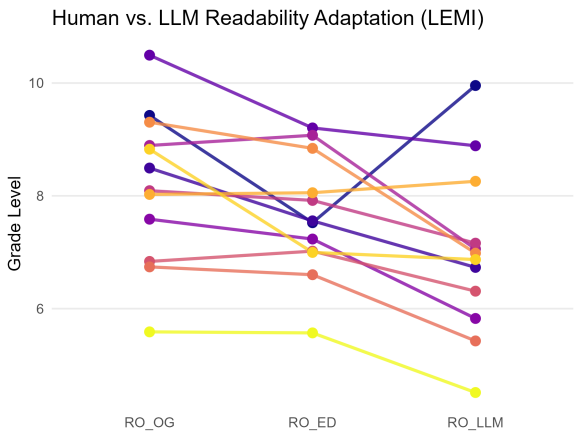

Figure 4: Grade-level comparison across published Romanian translations, human adaptations, and readability-constrained LLM translations

port this finding. Inflectional diversity, measured via the wordform–lemma ratio, is significantly higher in Romanian translations ($p = .0025$, $r = 1.00$), indicating greater morphological variation relative to English source texts.

Collectively, the grade-level indices, cross-linguistic readability scores, and morphological measures converge in showing a systematic translation-induced difficulty shift from English to Romanian in this corpus.

5.2. Readability-Constrained LLM Translation

We next evaluate whether readability-constrained LLM translation reduces difficulty relative to the published Romanian translations.

As illustrated in Figure 3, LLM translations (RO_LLM) consistently reduce grade-level estimates compared to the published Romanian versions (RO_OG). The LEMI-based index shows a median within-passage decrease of $\Delta = -1.46$ grade levels ($p = .0068$, $r = -0.90$), indicating a substantial downward shift in educational readability level.

Cross-linguistic readability patterns converge with this finding. LIX scores are significantly reduced in RO_LLM relative to *ro_og* ($p = .0033$, $r = -0.97$), reflecting shorter sentences and/or a lower proportion of long words.

At the morphological level, LLM translations exhibit significantly lower morphological feature density per token ($p = .0207$, $r = -0.77$), suggesting a measurable reduction in inflectional and agreement load. Lexical diversity, as measured by *MTLD*, is also significantly lower in RO_LLM ($p = .0167$, $r = -0.79$), consistent with vocabulary simplification.

Overall, these results indicate that readability-constrained LLM translation substantially attenuates translation-induced difficulty across structural, lexical, and morphological dimensions. The consistency of effect sizes across all 12 paired samples indicates a robust pattern of difficulty shifts, though larger-scale validation is required to confirm generalizability.

5.3. Human Adaptation vs. LLM Translation

We further compare human readability adaptations (RO_ED) to both the published translations and the readability-constrained LLM outputs.

Relative to the published Romanian translations (RO_OG), human adaptations yield a modest but statistically significant reduction in grade-level estimates (LEMI-based index; median $\Delta \approx -0.26$; $p = .045$, $r = -0.67$). *LIX* scores likewise decrease significantly ($p = .0053$, $r = -0.92$), indicating structural simplification.

However, the magnitude of reduction differs across adaptation types. LLM translations (*ro_llm*) exhibit a substantially larger median decrease relative to RO_OG (median $\Delta \approx -1.46$), as reported in the previous section. Direct comparison reveals that LLM translations yielded lower grade-level estimates than human adaptations (median difference ≈ 0.79 grades). While this difference fell just short of the conventional significance threshold ($p = .055$), the large effect size ($r = 0.64$) suggests a substantive divergence in simplification magnitude, likely limited by the statistical power of the small sample size ($N = 12$).

Figure 4 visualizes the trajectory across conditions. Most passages show a small decrease from RO_OG to RO_ED, followed by a larger decrease in RO_LLM. While individual passages vary, including one clear outlier, the overall pattern indicates that readability-constrained LLM translation tends

to reduce grade-level estimates more strongly than manual editing in this dataset.

Altogether, these findings suggest that controlled LLM translation not only approximates but, in this sample, often exceeds the magnitude of readability reduction achieved through human adaptation. At the same time, the marginal direct comparison ($p = .055$) indicates that differences between LLM and human editing should be interpreted cautiously.

5.4. Linguistic Mechanisms

To examine the linguistic mechanisms underlying the observed difficulty shifts, we analyze morphological and lexical complexity metrics.

At the morphological level, readability-constrained LLM translations (RO_LLM) exhibit a significant reduction in morphological feature density relative to the published translations (*ro_og*) ($p = .0207$, $r = -0.77$). This decrease suggests a measurable reduction in inflectional marking and agreement complexity in the LLM outputs. In contrast, human adaptations (*ro_ed*) do not show a statistically significant reduction in morphological feature density relative to *ro_og*, indicating that manual editing does not systematically reduce inflectional load in the same way.

Lexical diversity patterns show a similar asymmetry. *MTLD* decreases significantly in RO_LLM relative to RO_OG ($p = .0167$, $r = -0.79$), consistent with vocabulary simplification. Differences between RO_ED and RO_LLM on lexical diversity measures are limited and not statistically robust.

These findings suggest that readability-constrained LLM translation reduces difficulty not only through structural simplification (e.g., shorter sentences), but also through systematic reductions in morphological density and lexical diversity. In this dataset, the stronger grade-level reductions observed for RO_LLM appear to be accompanied by measurable changes in inflectional complexity, an effect particularly relevant for morphologically rich target languages such as Romanian.

5.5. Qualitative Analysis of Simplification Strategies

To better understand the linguistic mechanisms driving the observed quantitative shifts, we conducted a qualitative inspection of aligned sentence trios. This analysis reveals distinct simplification strategies between human editors and the readability-constrained LLM.

While human editors occasionally perform structural interventions, they often leave morphological and idiomatic complexities intact. Table 1 illustrates this contrast using an excerpt from *Inkling* (Code: `013-KO-PT`).

In this example, the human editor explicitly attempts to lower difficulty by splitting the long sentence into two shorter ones (*"...de umbre. Era cu ochii..."*). This effectively reduces sentence length, a key component of the LEMI and LIX metrics. However, the editor retains the *Perfect Simplu* tense (*se strecură*), a literary form often challenging for younger readers, and preserves the complex idiom *"cu ochii în patru"* (literally "with eyes in four").

The LLM, conversely, targets all three dimensions of difficulty. Structurally, it condenses the phrasing without splitting. Morphologically, it shifts from the literary *Perfect Simplu* to the standard conversational *Perfect Compus* (*a alunecat*), significantly lowering inflectional density. Lexically, it dissolves the idiom into a direct adjective (*atent* / "careful"). This demonstrates that while human editing in this corpus was primarily *remedial* (fixing length), the LLM generation was *transformative* (altering register and encoding).

6. Discussion

This study investigated translation-induced readability shifts and the effects of readability-constrained LLM translation in Romanian children's literature. The findings yield three main insights regarding the mechanics of automated simplification in morphologically rich languages.

First, published Romanian translations were consistently more difficult than their English originals. This pattern was confirmed across cross-linguistic (*LIX*), language-specific grade-level indices, and morphological variation measures. The convergence of structural and morphological indicators suggests that translation into a morphologically rich language may systematically increase surface complexity, even when semantic content is preserved. In this corpus, translation-induced difficulty appears to be driven not only by longer sentences or lexical choices, but also by increased inflectional diversity, which is a factor often overlooked in standard readability formulas.

Second, readability-constrained LLM translation substantially attenuated this difficulty shift. The median grade-level reduction relative to published translations was approximately 1.46 grade levels, accompanied by significant reductions in *LIX*, morphological feature density, and lexical diversity (*MTLD*). These results indicate that controlled LLM generation operates across multiple linguistic dimensions simultaneously. As observed in the qualitative analysis, the LLM did not merely shorten sentences; it actively normalized literary tenses (e.g., shifting from *perfect simplu* to *perfect compus*) and dissolved complex idiomatic constructions. This measurable decrease in morphological feature density suggests that LLM-based simplification can

Table 1: Comparison of Human vs. LLM adaptation strategies in passage `013-KO-PT`.

Condition	Text Segment	Key Features
English Original	"Along the shadowy hallway, he slid cautiously, keeping an eye out for Rickman, who also kept night hours."	Participial phrase (*keeping an eye out*); idiom; descriptive clause.
Published Translation	*"Se strecură prudent, de-a lungul holului plin de umbre, fiind cu ochii în patru să nu apară de undeva Rickman..."*	**Literary tense** (*Perfect Simplu*: *strecură*); **Complex idiom** (*ochii în patru*); **Syntactic expansion** (added sub-clause *să nu apară...*).
Human Adaptation	*"Se strecură atent, de-a lungul holului plin de umbre. Era cu ochii în patru să nu apară de undeva Rickman..."*	**Sentence splitting**; **Lexical edit** (*prudent* $\rightarrow$ *atent*); Retention of literary tense and idiom.
LLM Translation	*"A alunecat cu grijă pe holul întunecat, atent la Rickman, care era și el treaz noaptea."*	**Tense normalization** (*Perfect Compus*: *a alunecat*); **De-idiomatization** (*ochii în patru* $\rightarrow$ *atent*); **Syntactic compression**.

effectively target the specific inflectional burdens that characterize Romance languages.

Third, human adaptation and LLM simplification differed fundamentally in magnitude and strategy. Manual editing produced a modest median reduction (approximately 0.26 grade levels), whereas LLM outputs exhibited larger shifts. While the direct statistical comparison between human and LLM adaptation approached significance ($p = .055$), the substantial effect size ($r = 0.64$) points to a distinct divergence in approach. Human editors adhered to a *conservative* strategy, prioritizing fidelity and limiting interventions to local sentence splitting or lexical substitution. In contrast, the readability-constrained LLM adopted an *aggressive* structural strategy, rewriting entire syntactic architectures to meet the target grade level. The statistical margin likely reflects the small sample size ($N = 12$) rather than an absence of effect; the qualitative evidence confirms that the LLM performed types of simplification (morphological reduction) that human editors largely avoided.

These findings have implications for readability control in multilingual education. While the specific morphological effects observed here are tied to Romanian, the paired-corpus methodology and readability-controlled generation pipeline can be replicated for other language pairs. In languages such as Romanian, morphological complexity contributes substantially to processing load. Automated simplification systems that modulate inflectional density may therefore have a disproportionate impact on perceived difficulty compared to systems that only target sentence length. At the same time, the aggressive reduction of lexical diversity and morphological marking raises questions about the balance between accessibility and stylistic richness. While the LLM outputs were more readable, they occasionally sacrificed the literary register preserved by human editors.

From an NLP perspective, the results contribute empirical evidence that prompt-based control can achieve consistent difficulty reductions within a translation setting, unlike traditional post-hoc simplification pipelines. This suggests potential for integrated translation-and-adaptation workflows in educational publishing.

These findings have implications for readability control in multilingual education. While the specific morphological effects observed here are tied to Romanian, the paired-corpus methodology and readability-controlled generation pipeline can be replicated for other language pairs. In languages such as Romanian, morphological complexity contributes substantially to processing load. Automated simplification systems that modulate inflectional density may therefore have a disproportionate impact on perceived difficulty compared to systems that only target sentence length. At the same time, the aggressive reduction of lexical diversity and morphological marking raises questions about the balance between accessibility and stylistic richness. While the LLM outputs were more readable, they occasionally sacrificed the literary register preserved by human editors.

Several limitations must be acknowledged. The corpus is small ($N = 12$) and restricted to a single genre. While effect sizes are large and directionally consistent, replication on larger and more diverse datasets is necessary. Additionally, this study evaluates readability through automated indices; it does not include comprehension testing or human judgment of pedagogical appropriateness. Future work should incorporate reader-based validation to determine if the morphological reductions achieved by the LLM correspond to improved comprehension outcomes in young readers.

The findings demonstrate that translation-induced difficulty shifts are measurable in a morphologically rich target language, and that readability-

constrained LLM translation can mitigate these shifts through structural and morphological interventions that exceed the scope of standard human editing.

7. Conclusion

This study examined translation-induced readability shifts and the effects of readability-constrained LLM translation in Romanian children's literature. Using a paired four-condition design, we showed that published Romanian translations are consistently more difficult than their English originals across cross-linguistic (*LIX*), grade-level (FKGL/LEMI), and morphological measures, confirming that translation into a morphologically rich language can increase structural and inflectional complexity. Readability-constrained LLM translation substantially attenuated this shift, yielding a median reduction of approximately 1.46 grade levels relative to the published versions, alongside significant decreases in *LIX*, morphological feature density, and lexical diversity (*MTLD*). Human adaptation produced a smaller median reduction (approximately 0.26 grade levels). Although the direct comparison between LLM and human adaptation was marginal ($p = .055$, $r = 0.64$), LLM outputs generally achieved larger reductions. Methodologically, the study contributes a reproducible framework combining paired experimental design with morphologically informed readability metrics. These findings suggest that controllable LLM translation can meaningfully modulate readability, including inflectional complexity, in multilingual educational contexts.

8. Acknowledgments

We would like to thank **Quarticle Research SRL** for contributing to the technical integration of the automatic readability assessment interface within the LEMI platform (https://www.lemi.ro), which was developed by the authors. The functionality was employed for the automated computation of readability and linguistic complexity metrics in Romanian in the present study.

9. Bibliographical References

Mona Baker. 1993. Corpus linguistics and translation studies — implications and applications. In *Text and Technology*. John Benjamins.

Christian Bentz, Tatyana Ruzsics, Alexander Koplenig, and Tanja Samardžić. 2016. A comparison between morphological complexity measures: Typological data vs. language corpora. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 142–153, Osaka, Japan. The COLING 2016 Organizing Committee.

C.H. Björnsson. 1968. *Läsbarhet:*. Pedagogiskt Utvecklingsarbete vid Stockholms Skolor. 6. Liber; [Solna, Seelig].

Shoshana Blum-Kulka, Shoshana Blum-Kulka, and Julian House. 1996. Shifts of cohesion and coherence in translation.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Madalina Chitez, Mihai Dascalu, Aura Cristina Udrea, Cosmin Striletchi, Karla Csürös, Roxana Rogobete, and Alexandru Oravițan. 2024. Towards building the LEMI readability platform for children's literature in the Romanian language. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16450–16456, Torino, Italia. ELRA and ICCL.

Alina Maria Ciobanu, Liviu P. Dinu, and Flaviu Pepelea. 2015. Readability assessment of translated texts. In *Recent Advances in Natural Language Processing*.

Dana-Mihaela Cocargeanu. 2015. *Children's literature across space and time: the challenges of translating Beatrix Potter's tales into Romanian*. Ph.D. thesis, School of Applied Language and Intercultural Studies, Dublin City University.

Michael A. Covington and Joe D. McFall. 2010. Cutting the gordian knot: The moving-average type–token ratio (mattr). *Journal of Quantitative Linguistics*, 17(2):94–100.

Edward E Cureton. 1956. Rank-biserial correlation. *Psychometrika*, 21(3):287–290.

Mihai Dascalu, Philippe Dessus, Maryse Bianco, Stefan Trausan-Matu, and Jean-Luc Nespoulous. 2017. Readerbench: A multi-lingual framework for analyzing text complexity. *Educational Psychology Review*, 29:495–525.

Ion Giurgea. 2024. Article drop and case marking in romanian. *Isogloss J. Var. Roman. Iber. Lang.*, 10(2):1–23.

Mihai Iovanel. 2022. Children's literature in romania. *Transylvanian Review*, 31(1).

Ovidiu Ivancu. 2019. A pedagogical perspective on the definite and the indefinite article in the romanian language. challenges for foreign learners. *Verbum*, 10:1.

Kimmo Kettunen. 2014. Can type-token ratio be used to show morphological complexity of languages? *J. Quant. Linguist.*, 21(3):223–245.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.

Xiaofei Lu. 2012. The relationship of lexical richness to the quality of ESL learners' oral narratives. *Mod. Lang. J.*, 96(2):190–208.

Mounica Maddela, Fernando Alva-Manchego, and Wei Xu. 2021. Controllable text simplification with explicit paraphrasing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3536–3553, Online. Association for Computational Linguistics.

Louis Martin, Éric de la Clergerie, Benoît Sagot, and Antoine Bordes. 2020. Controllable sentence simplification. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4689–4698, Marseille, France. European Language Resources Association.

Philip M. McCarthy and Scott Jarvis. 2010. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42:381–392.

Mihaela Pirvulescu. 2002. Morphological paradigms and the role of tense. *Revue québécoise de linguistique*, 31(2):77–88.

Simina Sarbu. 2025. 12 cărți pentru copii premiate: Bestsellers în românia 2026. *Drimoland*.

SM Taheri and Gholamreza Hesamian. 2013. A generalization of the wilcoxon signed-rank test and its applications. *Statistical Papers*, 54(2):457–470.

Wouter van Loon, JJ Goeman, M Van Iterson, and M Fiocco. 2017. The power of the benjamini-hochberg procedure. *Leiden, The Netherlands: Leiden University*.

10. Appendices

10.1. LLM Prompting Protocol

The following base prompt was used to define the model's role and constraints. This instruction remained constant across all passages.

> You are a professional translator (English → Romanian) specialized in children's literature and readability control.
>
> Your task: translate the provided English text into Romanian while targeting a specified Romanian school grade level (e.g., grade 3–4). Maintain meaning, tone, and narrative voice.
>
> Hard constraints:
>
> - Do NOT add new content, explanations, or summaries.
> - Do NOT omit content.
> - Preserve character names and proper nouns exactly as in the source unless a standard Romanian form is conventional.
> - Preserve dialogue structure and paragraph breaks as much as Romanian norms allow.
> - Avoid "politeness inflation": do not introduce honorifics, formal pronouns (e.g., dumneavoastră), or extra politeness markers unless explicitly present in the source.
> - Keep register natural and age-appropriate.
>
> Readability controls (apply gently, without distorting meaning):
>
> - Prefer shorter sentences where possible; avoid heavy subordination.
> - Prefer common, high-frequency Romanian words; avoid rare/Latinate synonyms.
> - Avoid nominalizations and overly abstract phrasing when a simpler verb phrase works.
> - Keep cohesion clear (explicit subjects where needed, but don't over-repeat).
> - Keep the style literary/narrative (not textbook-like).
>
> Output format, no extra commentary, return ONLY:
>
> 1. TRANSLATION: <Romanian translation>

2. NOTES: bullet list of 3–6 concrete changes you made to target readability (e.g., “split a long sentence”, “replaced rare word X with common Y”).

For each passage, the base instruction was combined with the following variable prompt specifying the target readability level and the source text:

> Target readability: Romanian grade [X] (Romanian school system).
>
> Translate the text below from English into Romanian following the constraints.
>
> TEXT: « source passage »

The target grade level [X] was set to match the estimated readability level of the English source passage (FKGL), ensuring alignment between source difficulty and translation constraints.

All translations were generated using the GPT-5.2 model in a controlled, single-pass setup with deterministic decoding (low temperature), without iterative refinement or post-processing. All outputs were manually checked to ensure content fidelity, including the absence of omissions, additions, or hallucinated content. In addition to the translation, the model produced a short list of 3–6 notes describing the simplification strategies applied (e.g., sentence splitting, lexical simplification). These notes were not included in the quantitative analysis but were used to support qualitative inspection of simplification strategies (see Section 5.5).

10.2. Source Texts

The full dataset is avaiable on *HuggingFace* via https://huggingface.co/datasets/karlacsuros/lemitranslations

Table 2: Source texts included in the corpus

Code	Title	Author	Year
011-RG-DT	My Father’s Dragon	Ruth Stiles Gannett	1948
013-KO-PT	Inkling	Kenneth Oppel	2018
018-RT-CS	Who in the World is Carmen Sandiego	Rebecca Tinker	1998
019-PG-VC	Treasure Hunters	James Patterson	2013
020-JK-JP	Diary of a Wimpy Kid: Diper Överlöde	Jeff Kinney	2022
022-JB-DT	The Terrible Two	Jory John & Mac Barnett	2015
023-AD-NK	Ninja Kid #1: From Nerd to Ninja	Anh Do	2019
028-LP-TG	The Brilliant World of Tom Gates	Liz Pichon	2011
029-DW-MI	Awful Auntie	David Walliams	2014
057-JB-CB	Little Book of Bob	James Bowen	2018
060-EC-MR	Mariella, Queen of the Skies	Eoin Colfer	2018
061-SS-FF	Mighty Murphy	Shelley Swanson Sateren	2016